Soren Nelson

Data Mining CS 6140

**Intermediate Report**

**Data Collection:**

We started our data collection by getting the most popular political hashtags. Using the Tweepy api we gathered 48 for the February 7th 2020 Democratic debate and 89 hashtags for the Democratic debate on February 19th 2020. We then used these hashtags as search terms to get all tweets using these hashtags with more than 5 favorites.

**Polarity and Subjectivity Per Candidate**

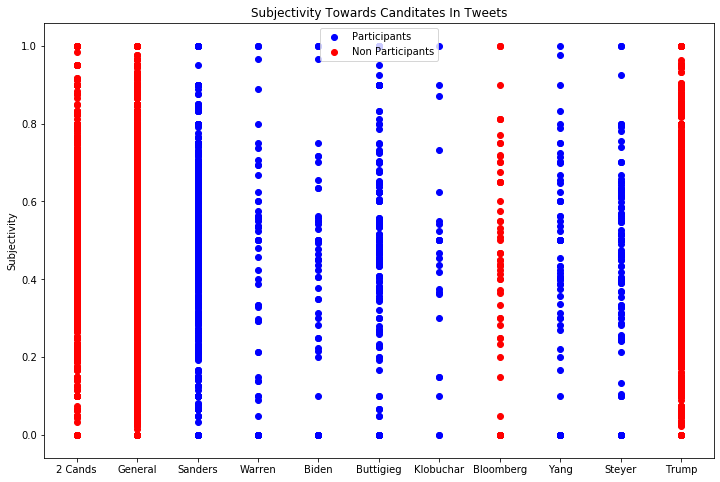
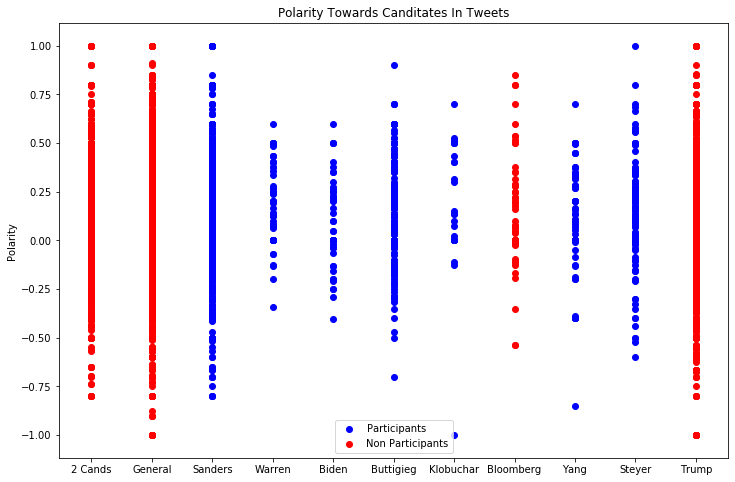
We wanted to look at what people are saying on twitter about a specific candidate and if that was correlated to what the candidate said during the debates as well as how many votes the candidate got. To find the subject of a given tweet we checked whether a tweet used a single candidate's name. We separated out all tweets that referred to more than one candidate and any tweet that didn’t mention a candidate. This left us with the following distributions:

Two Candidates: 1326, No candidates: 5600, Sanders: 978, Warren: 60, Biden: 60, Buttigieg: 226, Klobuchar: 26, Bloomberg: 67, Yang: 79, Steyer: 152, Trump: 2211 for the first debate,

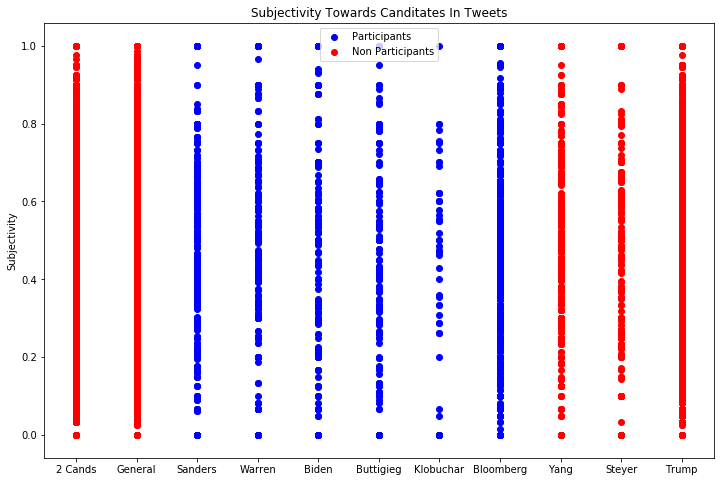
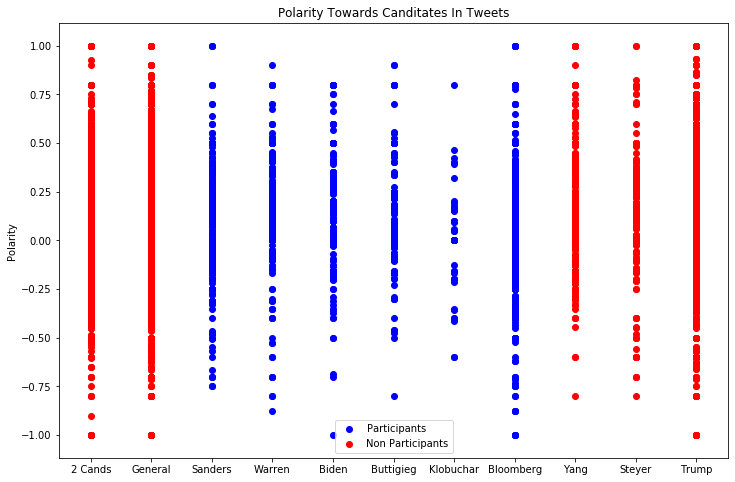
Two Candidates: 2966, No Candidates: 7283, Sanders: 405, Warren: 315, Biden: 272, Buttigieg: 162, Klobuchar: 67, Bloomberg: 803, Yang: 489, Steyer: 201, Trump: 2689 for the second debate.

President Trump had the most tweets about him even though he was not participating in either debate. Interestingly, both Yang and Steyer had more tweets about them in the second debate even though they both only participated in the first. There were more tweets we gathered on the second debate night which is most likely due to the February 20th coming much closer to the Primary’s.

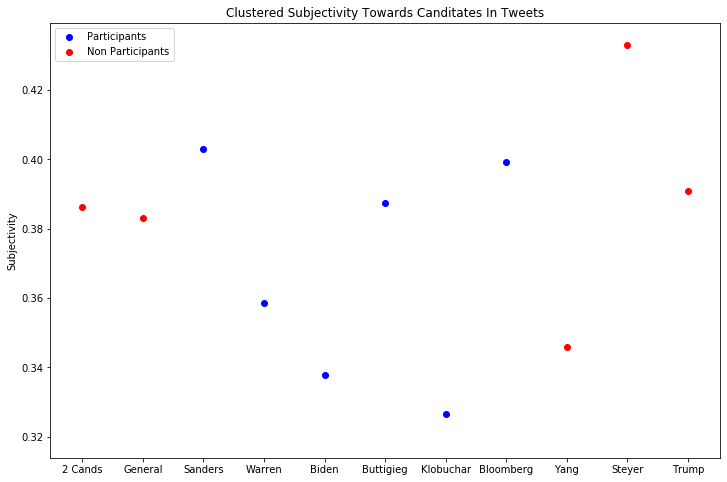
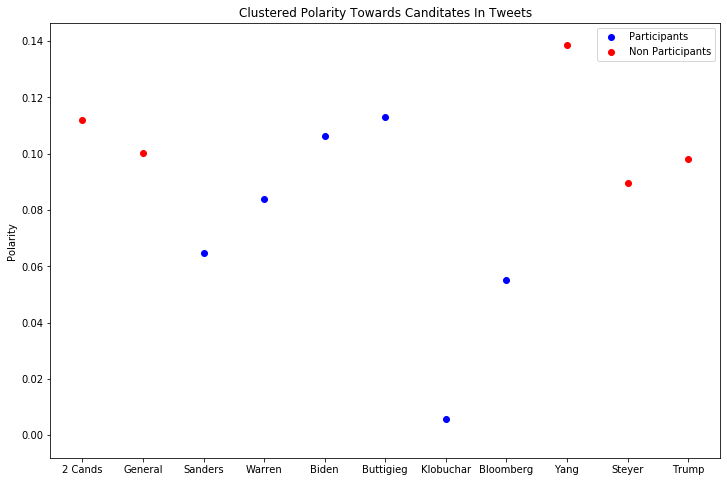
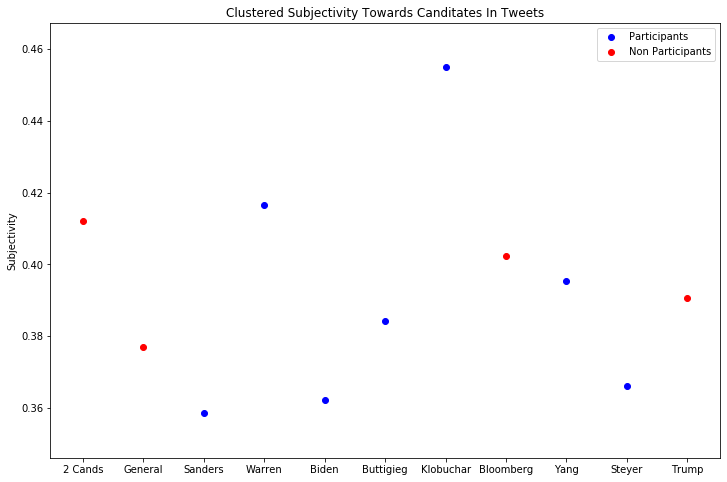
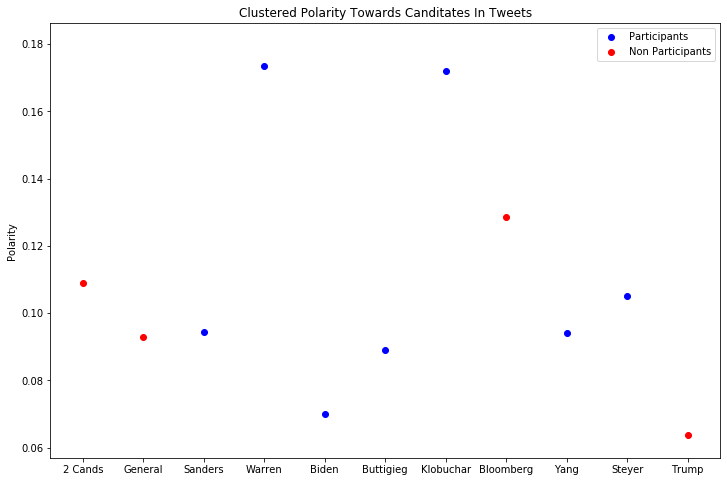
We took all these tweets and, using Python’s textblob, we gathered the polarity and subjectivity of the tweets about each candidate. The following are plots from the first debate:

****

and ones from the second:



We can see that the majority of the candidates lean slightly more positive and more objective than subjective. We felt this wasn’t very informative and difficult to gather much data so using this data we ran Lloyd’s Algorithm to cluster the data. We initialized a cluster at 0.0 for each candidate on the polarity data and 0.5 on the subjectivity data. After clustering we get the following plots:



We can then clearly see that all candidates are talked about fairly neutral although more positive than negative. We can also see who did well and who didn’t in the debate, such as Klobuchar being extremely positive in the first debate and the most neutral in the second debate. After gathering data for another debate we intend on plotting each of these debates together to further extrapolate trends. We will also be analyzing the debate transcripts in the same manner which may lead to more insights.